Immunoinformatics: Current trends and future directions

Joo Chuan Tong[1] and Ee Chee Ren[2,3]

[1] Institute for Infocomm Research, 1 Fusionopolis Way, #21-01 Connexis, South Tower, Singapore 138632, Singapore
[2] Department of Microbiology, Faculty of Medicine, National University of Singapore, Singapore 117597, Singapore
[3] Singapore Immunology Network, 20 Biopolis Way, #03-01 Immunos, Singapore 138668, Singapore

Immunoinformatics has recently emerged as a critical field for accelerating immunology research. Although still an evolving process, computational models now play instrumental roles, not only in directing the selection of key experiments, but also in the formulation of new testable hypotheses through detailed analysis of complex immunologic data that could not be achieved using traditional approaches alone. Immunomics, which combines traditional immunology with computer science, mathematics, chemistry, biochemistry, genomics and proteomics for the large-scale analysis of immune system function, offers new opportunities for future bench-to-bedside research. In this article, we review the latest trends and future directions of the field.

It has long been recognized that computational methods and resources has the potential to accelerate immunology research, but recent advances in genomic and proteomic technologies has radically transformed the opportunities. Sequencing of the human and other model organism genomes has produced increasingly large volumes of data relevant to immunology research. At the same time, huge amounts of functional, clinical and epidemiologic data are being reported in the scientific literature and deposited in various specialist repositories and clinical records. Together, this accumulated information reflects the current state of knowledge on human immunology and disease, and represents a goldmine for researchers looking for insights into the mechanisms of immune function and disease pathogenesis.

The need to handle this rapidly growing immunological resource has given rise to the field known as immunoinformatics. Immunoinformatics, otherwise known as computational immunology, is now an essential component of modern immunology research. It lies at the interface between computer science and experimental immunology, and represents the use of computational methods and resources for the understanding, generation, processing, and propagation of immunological information. Historically, immunoinformatics began over 90 years ago with the theoretic modeling of malaria epidemiology [1]. At that time, the emphasis was on the use of mathematics to guide the study of disease transmission. Since then, the field has expanded to cover all other aspects of immune system processes and diseases. Immunomics, which combines traditional immunology with computer science, mathematics, chemistry, biochemistry, genomics and proteomics for large-scale analysis of immune system functions, offers new opportunities for future bench-to-bedside research. Computational systems immunology, which aims to study the complex protein–protein interactions and networks, allows a better understanding of immune responses and their role during normal, disease and reconstitution states. Syndromic surveillance, which focuses on monitoring disease trends using health-related data, allows early detection and response to potential outbreaks. This article provides an introduction to the growing literature in this area, with special emphasis on the latest trends and future directions.

The combinatorial immune system

The human immune system is highly combinatorial in nature. A large repertoire of immunoglobulins (IG) ($\sim 10^{12}$) and T cell receptors (TR) ($\sim 10^{12}$) is generated by mechanisms such as the combinatorial diversity of the variable (V), diversity (D) and joining (J) genes, the N-diversity, and for IG, the somatic hypermutations

Corresponding author: Tong, J.C. (jctong@i2r.a-star.edu.sg)

1359-6446/06/$ - see front matter © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2009.04.001

[2,3] The human leukocyte antigen (HLA) haplotype influences individual immune responses against specific antigens. As of January 2009, 3304 alleles had been identified and deposited in the IMGT/HLA Database [4]. Since a fully heterogeneous individual may possess up to six different HLA class I alleles and an equal number of class II alleles, the theoretical number of HLA haplotypes is greater than 10^{12}. Binding studies showed that HLA class I binding peptides and the core of class II binding peptides are

TABLE 1

Some existing immunological databases reviewed in this article.

Database	Summary	URL	Refs
IMGT®	IMGT® is an integrated knowledge resource specialized in IG, TR, MHC, IG superfamily, MHC superfamily and related proteins of the immune system of human and other vertebrate species. IMGT® comprises 6 databases, 15 on-line tools for sequence, gene and 3D structure analysis, and more than 10,000 pages of resources Web. Data standardization, based on IMGT-ONTOLOGY, has been approved by WHO/IUIS	http://www.imgt.org/	[11]
SYFPEITHI	SYFPEITHI is a searchable database that contains more than 4500 records of MHC ligands and peptide motifs from humans and other species. Hyperlinks to the EMBL and PubMed databases are included. In addition, ligand predictions are also available for a number of MHC allelic products	http://www.syfpeithi.de/	[5]
IEDB	IEDB is a resource center for data related to antibody and T cell epitopes for humans, non-human primates, rodents, and other animal species. T and B cell epitope prediction tools, epitope analysis tools, as well as peptide processing predictions are also available. As of January 2009, IEDB stores 5138 references, 115,906 records, 80,884 distinct structures, and 38,580 distinct epitopes	http://www.immuneepitope.org/	[10]
AntiJen	AntiJen contains over 24,000 entries of binding data on MHC ligands, TR-peptide-MHC complexes, T cell epitopes, TAP, B cell epitopes and immunological protein–protein interactions. Peptide library, copy numbers and diffusion coefficient data are also included	http://www.jenner.ac.uk/antijen/	[8]
MHCBN	MHCBN contains information on 20,717 MHC binders, 4022 MHC non-binders, 1053 TAP binders and non-binders, and 6722 T cell epitopes	http://www.imtech.res.in/raghava/mhcbn/	[7]
Los Alamos HIV databases	The Los Alamos HIV databases contain data on HIV sequences, 905 HIV-1 cytotoxic T cell epitopes, 1023 HIV-1 helper T cell epitopes, 1448 antibody-binding sites, drug resistance-associated mutations, and vaccine trials. The website also provides access to a large number of tools that can be used to analyze these data	http://www.hiv.lanl.gov/	
IPD	IPD is a set of specialist databases related to the study of polymorphic genes in the immune system. It consists of four databases: IPD-KIR for sequences of killer-cell IG-like receptors; IPD-MHC for MHC sequences of different species; IPD-HPA for alloantigens expressed only on platelets; and IPD-ESTAB for access to the European Searchable Tumour Cell-Line Database, a cell bank of immunologically characterized melanoma cell lines	http://www.ebi.ac.uk/ipd/	[15]
Epitome	Epitome stores information of 142 antigens from protein–antibody complex structures	http://www.cubic.bioc.columbia.edu/services/Epitome/	[9]
Allergen Nomenclature Database	The Allergen Nomenclature Database contains information of allergens and isoallergens developed and maintained by the Allergen Nomenclature sub-committee of the IUIS. Data submissions are accepted and annotated by the committee members	http://www.allergen.org/	[17]
BIFS	BIFS contains information on 453 food allergens (64 animals, 389 plants), 645 non-food allergens, and 75 wheat gluten proteins	http://www.iit.edu/~sgendel/fa.htm	[16]
SDAP	SDAP stores information of 887 allergenic proteins. It contains various tools for FAO/WHO allergenicity tests and assessing the IgE-binding potential of genetically modified food proteins	http://www.fermi.utmb.edu/SDAP/	[18]
FARRP	FARRP contains 1251 sequences of known and putative allergens derived from scientific literature and public databases	http://www.farrp.org/	[19]
Allergome	Allergome emphasizes the annotation of allergens that cause IgE-mediated disease. The database contains information derived from 5800 selected scientific literatures	http://www.allergome.org/	[20]

predominantly nine amino acids long. Consequently, the number of potential nonameric peptide candidates easily exceeds 10^{11}. The astronomically high diversity of the immune system components, and also the complexity of its regulatory pathways, demands new approaches for accelerating immunology research. An immunoinformatics strategy, which combines database design, mathematical modeling and high-performance computing, provides a highly parallel, rational solution and works effectively on a massive scale.

Immunological databases

Sequencing of the human and other model organism genomes has produced increasingly large amounts of data relevant to the study of human immune systems and disease. A total of 27 immunological databases are currently (January 2009) described in the Nucleic Acids Research Molecular Biology Database Collection (http://www3.oup.co.uk/nar/database/c/). Some of these databases are reviewed below (Table 1). Immune epitope databases [5–10] are useful for major histocompatibility complex (MHC), TR and IG-binding analysis, with direct implications in component vaccine design and analysis of host–pathogen interactions. Important sources of MHC ligand data include the SYFPEITHI database [5], which contains >4500 records of MHC ligand data and peptide motifs; the HIV Molecular Immunology Database [7], which stores information on 905 HIV-1 cytotoxic T cell epitopes, 1023 HIV-1 helper T cell epitopes and 1448 antibody-binding sites and the Immune Epitope Database and Analysis Resource (IEDB) [10], which records data related to IG and T cell epitopes for humans, non-human primates, rodents, and other animal species.

Immune sequence databases are valuable for research in autoimmune disorders, infectious diseases, cancer, immunotherapy and immunoprophylaxis. IMGT®, the international ImMunoGeneTics information system® (http://imgt.org), serves as a central resource for IG, TR, MHC, and related proteins of the immune system of human and other vertebrates [11]. IMGT contains five databases: (1) IMGT/LIGM-DB [12] with 132,746 IG and TR sequences from human and 232 vertebrate species; (2) IMGT/MHC-DB with sequences of 2292 HLA class I alleles, 1,012 HLA class II alleles, and 106 non-HLA alleles; (3) IMGT/GENE-DB [13] with 1922 genes and 2988 alleles of human, mouse, rat and rabbit IG and TR genes; (4) IMGT/PRIMER-DB with 1864 primer records of IG and TR from 11 species; and (5) IMGT/3Dstructure-DB [14] with 1562 records of IG, TR, and MHC proteins with known 3D structures. The Immuno Polymorphism Database (IPD) [15] consists of four specialist databases: (1) IPD-KIR contains the allelic sequences of 233 killer-cell immunoglobulin-like receptors; (2) IPD-MHC details the MHC sequences of a number of different species; (3) IPD-HPA stores data which define the human platelet antigens; (4) IPD-ESTAB provides access to the European Searchable Tumour Cell-Line Database (ESTDAB), a cell bank of immunologically characterized melanoma cell lines.

Online resources for allergy information are also available. Such data is valuable for investigation of cross-reactivity between known allergens and analysis of potential allergenicity in proteins. The Biotechnology Information for Food Safety (BIFS) database [16] contains information on 453 food allergens (64 animals, 389 plants), 645 non-food allergens, and 75 wheat gluten proteins. The Allergen Nomenclature database of the International Union of Immunological Societies (IUIS) (http://www.allergen.org) provides a centralized system to ensure uniformity and consistency of allergen designations [17]. As of January 2009, more than 779 allergens and isoallergens from over 150 different species that can induce IgE-mediated allergy (reactivity >5%) in humans are described. The Structural Database of Allergen Proteins (SDAP) [18] stores information of 887 allergenic proteins. The Food Allergy Research and Resource Program (FARRP) Protein Allergen-Online Database [19] contains 1251 sequences of known and putative allergens derived from scientific literature and public databases. Allergome [20] emphasizes the annotation of allergens that result in an IgE-mediated disease. The database currently (last update in November 2004) contains information derived from 5800 selected scientific literatures. Web sites of interest for immunologists are listed in [21] and links to databases, tools and resources in immunoinformatics are available in 'The IMGT immunoinformatics page' at http://imgt.org.

Computational tools

A wide variety of computational, mathematical and statistical methods has been reported in the literature, ranging from text mining, information management, sequence analysis, molecular interactions, to advanced systems simulation. Text mining of biomedical literature is at its formative stages. Attempts are being made for the extraction of interesting and complex patterns from non-structured text documents in the immunological domain. Examples include categorization of allergen cross-reactivity information [22], identification of cancer-associated gene variants [23], and the classification of immune epitopes [24].

Conventional sequence analysis tools, such as ClustalW, BLAST, and TreeView, as well as specialized immunoinformatics tools, such as IMGT/V-QUEST [25] for IG and TR sequence analysis, IMGT/Collier-de-Perles [26] and IMGT/StructuralQuery [14] for IG variable domain structure analysis, allow the inference of functional, structural, or evolutionary relationships between DNA or protein sequences. Methods that rely on sequence comparison are diverse and have been applied to analyze HLA sequence conservation [4], help verify the origins of human immunodeficiency virus (HIV) sequences [27], and construct homology models for the analysis of hepatitis B virus polymerase resistance to lamivudine and emtricitabine [28]. Computational models that focus on protein–protein interactions and networks include procedures for T and B cell epitope mapping, proteasomal cleavage site prediction, and TAP–peptide prediction [29]. The availability of experimental data is a necessity for developing efficient and robust machine-learning models to predict various molecular targets. Methods based on structure-guided design are likely to become very powerful in the years to come, with the rapid increase in structural data generated by molecular biology initiatives. Cellular automata that utilize sophisticated mathematical formulae to describe a wide range of complex virus–host relations were also reported. Specific examples include simulating cognate recognition and response in the immune system [30], modeling B cell maturation [31], and analyzing MHC polymorphism under host–pathogen co-evolution [32].

Computational vaccinology

Vaccination is widely regarded as one of the most successful public health intervention measures in the fight against infectious diseases, allergies, neurodegenerative diseases, autoimmune disorders and some cancers. A successful example was the worldwide eradication of smallpox. The aim of vaccination is to prime the immune system in order to generate immunological memory so that a heightened immune response will be mounted upon exposure to the specified pathogen. Vaccines may be live attenuated whole organisms, killed micro-organisms, subunit antigens or toxoids. Vaccines based on killed pathogens may fail if the pathogen is denatured by extreme heat or chemical reactions. While it is possible to design attenuated vaccines using weakened forms of pathogens, subunit or peptide-based vaccines is generally preferred to reduce the risk of adverse reaction such as clinical manifestation of disease. In the past decade, computational methods for mapping immunogenic epitopes have been actively developed to facilitate the discovery of suitable vaccine candidates. An overview of these methods is described in this section.

B cell epitope prediction

B cell epitopes are antigenic determinants on the surface of pathogens that interact with B cell receptors. The B cell receptor (BCR)-binding site is primarily hydrophobic, consisting of six hypervariable loops of variable length and amino acid composition. B cell epitopes can be either continuous or discontinuous. Approximately 10% of B cell epitopes are contiguous, consisting of a linear stretch of amino acids along the polypeptide chain. Most B cell epitopes, though, are non-contiguous in nature, where distant residues are brought into spatial proximity by protein folding. It has been reported that not all residues within an epitope are functionally important for binding, and the specificity could be reduced or eliminated by single-site amino acid substitution. The astronomical combinatorial space of molecular interactions calls for the use of computational tools and mathematical models for systematic screening and analysis.

A variety of methods for the modeling and prediction of B cell epitopes have been reported. One of the biggest challenges in this field is the lack of standards in defining B cell epitopes which has a direct impact on the selection and development of appropriate tools for analysis [33]. Current efforts are primarily focused on the design of predictors for continuous epitopes, in part, due to lower complexity in system development, and also because the experimental design of conformational epitopes is non-trivial. Propensity scales are commonly used to guide the mapping of B cell epitopes. Hopp and Woods [34] introduced the use of hydrophilicity scales for locating protein antigenic determinants. Pellequer *et al.* [35] applied the use of turn propensities for the analysis on 85 continuous epitopes in 14 proteins. Other amino acid propensities were also proposed, but the effectiveness of such an approach has, so far, been controversial. A benchmark on the performance of 484 amino acid propensity scales for B cell epitope prediction revealed that the best set of scales and parameters performed only slightly better than random [36]. Machine learning techniques have been applied, but also achieved limited success for predicting continuous epitopes [37]. On the contrary, structure-based predictors for B cell epitope mapping are gaining dominance, due to (1) the rapidly increasing number of three-dimensional (3D) structures of antibody–antigen complexes available in the PDB and in IMGT/3Dstructure-DB, and (2) the ability to predict both continuous and discontinuous epitopes. In fact, the use of structural data parallels the noticeable shift in B cell epitope prediction methodologies over the past few years, away from sequence-derived properties to much more structure-guided designs [38].

T cell epitope prediction

T cells recognize antigens as short peptide fragments in association with MHC molecules on antigen-presenting cells. Two classes of T cells are available: (1) CD8+ T cytotoxic (Tc) cells, which recognize peptides displayed by MHC class I molecules, and (2) CD4+ T helper (Th) cells, which recognize peptides in association with MHC class II molecules. Tc cells release cytotoxins which are responsible for cell lysis, and granzymes which induces apoptosis. Th1 cells produce interferon γ (IFN-γ) and tumor necrosis factor β (TNF-β) and are involved in delayed-type hypersensitivity (DTH) reactions. In contrast, Th2 cells produce interleukin 4 (IL-4), IL-5, IL-10 and IL-13, which are responsible for strong antibody responses, including the activation and recruitment of IgE antibody-producing B cells, mast cells, eosinophils, and the inhibition of several macrophage functions.

Computational methods for predicting T cell epitopes and MHC-binding peptides have been extensively explored and reviewed elsewhere [39]. These include procedures based on binding motifs, binding matrices, decision trees, hidden Markov models (HMM), support vector machines (SVM), artificial neural networks (ANN), quantitative structure–activity relationship (QSAR) analysis, homology modeling, protein threading and docking techniques. In the last decade, much emphasis has been placed on the design of computational technologies that allow the prediction of promiscuous peptides capable of binding to a wide array of MHC molecules [40]. This approach allows the design of peptide vaccines with improved global coverage by ensuring that HLA alleles that are present in most individuals from all major ethnic groups may bind to at least one of the peptides in the vaccine. Tools for predicting MHC class I binding in higher vertebrates have also been reported [41]. Dynamic activities over the past 2 years have also seen at least six reports of algorithms that attempt to simulate the cell-mediated immune system by integrating the different sub-components of the antigen processing and presentation pathway such as TAP, proteasome, and MHC [42,43]. These tools are particularly useful for screening large sets of protein antigens, such as those encoded by complete viral genomes.

Allergy informatics

Current efforts in allergy informatics are primarily focused on quality data management, T and B cell epitope prediction, as well as the assessment of allergenicity and allergic cross-reactivity. Standards for assessing protein allergenicity are still in their formative stages. The World Health Organization (WHO) and Food and Agriculture Organization (FAO) proposed guidelines for evaluating allergenicity of genetically modified foods. According to the Codex alimentarius, a protein is potentially allergenic if it possesses an identity of ≥6 contiguous amino acids or ≥35% sequence similarity over an 80 amino acid window with a known allergen. Although these recommendations are in place, their inherent limitations are starting to become apparent and exceptions to the rules have been well reported [44]. Computational algorithms have been actively developed to help assess the allergenic potential of genetically modified food crops, bio-pharmaceuticals and various other products on the consumer market [45]. An example was the use of the FASTA3 algorithm with k-nearest-neighbour (kNN) classifier for assessing potential food allergenicity of newly introduced proteins [46]. The use of Fourier transform to detect compact patterns in allergens was also reported [44]. Implementation and controlled comparisons of allergen prediction systems is difficult, however, due to poorly defined standards for identifying allergenicity and the lack of experimental non-allergenic sequences. Many new allergen prediction systems use hypothetical or inferred non-allergenic sequences for system training and testing. At present, information is limited about the usefulness of much of these systems, and significant costs may be incurred in implementing, improving and managing these systems, as well as investigating false alarms. As more experimental validations are performed, the relative value of the different approaches will be known and it should be expected that the current FAO/WHO recommendations will also be refined.

Infectious disease informatics

There is intense interest in the use of informatics for the surveillance, modeling and response to infectious diseases. Syndromic surveillance has taken advantage of new technologies to model how diseases, especially unexpected emerging infections such as pandemic influenza or severe acute respiratory syndrome (SARS), could spread through a community. Mathematical modeling, pattern recognition and aberration detection are useful for screening data to identify patterns that warrant further public health investigation, to enhance recognition of disease outbreak patterns, and allow the monitoring and evaluation of control strategies. The Centers for Disease Control and Prevention framework on public health surveillance systems [47] highlighted the importance of data quality, consistency and accessibility, and directed particular attention to the measurement of timeliness and validity for outbreak detection as well as improved coordination and sharing of information. Specific examples of syndromic surveillance systems include autoregressive modeling of influenza-like illnesses in Minnesota [48], and Digital Ring Fence (DRiF) strategy for the containment of acute infectious outbreak [49].

There have been remarkable advances in deciphering human immune responses to various pathogens by integrating genomics and proteomics with bioinformatic strategies that allow a rational approach to target validation. Many exciting developments in large-scale screening of pathogens are currently taking place, including attempts for systematic mapping of B and T cell epitopes of National Institute of Allergy and Infectious Diseases (NIAID) category A-C pathogens. These pathogens include *Bacillus anthracis* (anthrax), *Clostridium botulinum* toxin (botulism), *Variola major* (smallpox), *Francisella tularensis* (tularemia), viral hemorrhagic fevers, *Burkholderia pseudomallei*, *Staphylococcus enterotoxin* B, yellow fever, influenza, rabies, Chikungunya virus, among others. Rule-based systems have been reported for the automated extraction and curation of influenza A records [50]. Information theory has been used to measure the variability of influenza A virus proteome [51]. This, combined with T cell epitope prediction algorithms, allows the identification of conserved regions in viral sequences that are prime targets for epitope-based T cell vaccine formulations. Attempts are also being made to describe the relationship between T cell epitope antigenic diversity and protein sequence diversity of dengue virus [52] and escape mutations in HIV-1 gag [53], which are of direct importance for peptide vaccine formulation. In the next few years, increased integration of surveillance technologies with sequence analysis and antigenicity assessment tools will allow detailed analysis of the future disease evolution patterns, and play a more prominent role in infection control and health care epidemiology.

Cancer informatics

Cancer progression is a form of somatic evolution in which certain mutations provide cancer cells with a selective growth advantage. A number of cancer genome projects are currently underway to identify novel mutations that drive tumorigenesis. An example of a targeted approach for assessing mutations and cancer risk has been reported by Kaminker *et al.* [54], in which the algorithm CanPredict was used to indicate how closely a specific gene resembles known cancer-causing genes. Protein–protein interaction networks provide valuable information on tumorigenesis in humans. Hernández *et al.* [55] has recently studied the coordinated function of cancer proteins in the human interactome. This work gives a good indication that cancer proteins are central to information exchange and propagation, and are specifically organized to promote tumorigenesis. Jonsson and Bates [56] performed a detailed analysis of cancer proteins in a human interactome constructed by computational methods, and showed that cancer proteins exhibit a network topology that is different from normal proteins in the human interactome. A number of computational models for classifying cancer subtypes based on epigenetic marks have also been reported. Specific examples include the use of SVMs for discriminating between acute lymphoblastic leukemia and acute myeloid leukemia [57], as well as the use of Manhattan distance and average linkage algorithms for hierarchical cluster analysis of human colorectal tumors [58]. Cancer epigenetics will be a major growth area in this domain, with the maturation of key initiatives such as the European Union (EU) funded CancerDip Consortium.

Conclusion

Realizing the full benefits of the informatics revolution will require significant advances in the efficiency with which new data is discovered, processed, interpreted and made accessible to researchers. The next few years will see increased interest in the use of cluster computing and distributed systems for large-scale data analysis and screening. With the explosion in the number, variety and sophistication of resources and analysis tools, the challenge is to integrate the strengths and not the weaknesses of each approach. Computational algorithms that model different aspects of the human immune system and disease have been described [42,43]. On the other hand, cellular automata have also been proposed for exploring a wide range of virus–host relations [30]. The different bioinformatic and mathematical modeling approaches, in combination with advances in computational infrastructures, allow the construction of new models that are orders of magnitudes more complex than those currently available and facilitate a system-level understanding of the structure and dynamics of cellular and organism functions. Already, a number of 'system biology' approaches for immunological studies, such as the EU funded ImmunoGrid project [59] and the National Cancer

Institute's Integrative Cancer Biology Program [60] have been reported. With the increasing availability of correlative data and maturation of key technologies for systems biology, it should be expected that many more such system-driven approaches will be developed that could lead to improved understanding of development and homeostasis of immune system processes.

References

1 Ross, R. (1916) An application of the theory of probabilities to the study of *a priori* pathometry. Part I. *Proc. R. Soc. Lond. Ser. A* 92, 204–230

2 Lefranc, M.-P. and Lefranc, G. (2001) *The Immunoglobulin FactsBook*. Academic Press, London, UK 458 pp.

3 Lefranc, M.-P. and Lefranc, G. (2001) *The T Cell Receptor FactsBook*. Academic Press, London, UK 398 pp.

4 Robinson, J. *et al.* (2001) IMGT/HLA Database – a sequence database for the human major histocompatibility complex. *Nucleic Acids Res.* 29, 210–213

5 Rammensee, H. *et al.* (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50, 213–219

6 Korber, B.T.M. *et al.*, eds. (2006/2007) *HIV Molecular Immunology*, Los Alamos National Laboratory, Theoretical Biology and Biophysics, Los Alamos, New Mexico

7 Bhasin, M. *et al.* (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19, 665–666

8 Toseland, C.P. *et al.* (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res.* 1, 4

9 Schlessinger, A. *et al.* (2006) Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res.* 34, D777–D780

10 Peters, B. *et al.* (2005) The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol.* 3, e91

11 Lefranc, M.P. *et al.* (2005) IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* 37, D1006–D1012

12 Giudicelli, V. *et al.* (2006) IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.* 34, D781–784

13 Giudicelli, V. *et al.* (2005) IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* 33, D256–261

14 Kaas, Q. *et al.* (2004) IMGT/3D structure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res.* 32, D208–D210

15 Robinson, J. *et al.* (2005) IPD – the Immuno Polymorphism Database. *Nucleic Acids Res.* 331, D523–D526

16 Gendel, S.M. (1998) Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food Nutr. Res.* 42, 63–92

17 Hoffman, D. *et al.* (1994) Allergen nomenclature. *Bull. World Health Organ.* 72, 796–806

18 Ivanciuc, O. *et al.* (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.* 31, 359–362

19 Hileman, R.E. *et al.* (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.* 128, 280–291

20 Mari, A. *et al.* (2005) Allergome – a database of Allergenic molecules: structure and data implementations of a web-based resource. *J. Allergy Clin. Immunol.* 115, S87

21 Lefranc, M.P. (2006) Web sites of interest to immunologists. *Curr. Protoc. Immunol.* Appendix 1, Appendix 1J

22 Miotto, O. *et al.* (2005) Supporting the curation of biological databases with reusable text mining. *Genome Inform.* 16, 32–44

23 McDonald, R. *et al.* (2006) An automated procedure to identify biomedical articles that contain cancer-associated gene variants. *Hum. Mutat.* 27, 957–964

24 Wang, P. *et al.* (2007) Automating document classification for the Immune Epitope Database. *BMC Bioinform.* 8, 269

25 Brochet, X. *et al.* (2008) IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res.* 36, W503–508

26 Kaas, Q. *et al.* (2007) IG, TR and IgSF, MHC and MhcSF: what do we learn from the IMGT Colliers de Perles? *Brief. Funct. Genomic Proteom.* 6, 253–264

27 Deng, W. *et al.* (2007) ViroBLAST: a stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. *Bioinformatics* 23, 2334–2336

28 Das, K. *et al.* (2001) Molecular modeling and biochemical characterization reveal the mechanism of hepatitis B virus polymerase resistance to lamivudine (3TC) and emtricitabine (FTC). *J. Virol.* 75, 4771–4779

29 Montañez, R. *et al.* (2006) Information integration of protein–protein interactions as essential tools for Immunomics. *Cell. Immunol.* 244, 84–86

30 Seiden, P.E. and Celada, F. (1992) A model for simulating cognate recognition and response in the immune system. *J. Theor. Biol.* 158, 329–357

31 Shahaf, G. *et al.* (2004) Screening alternative models for transitional B-cell maturation. *Int. Immunol.* 16, 1081–1090

32 Borghans, J.A.M. *et al.* (2004) MHC polymorphism under host–pathogen coevolution. *Immunogenetics* 55, 732–739

33 Greenbaum, J.A. *et al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recogn.* 20, 75–82

34 Hopp, T.P. and Woods, K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* 78, 3824–3828

35 Pellequer, J. *et al.* (1991) Predicting the location of continuous epitopes in proteins from their primary structure. *Methods Enzymol.* 203, 176–201

36 Blythe, M.J. and Flower, D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* 14, 246–248

37 Saha, S. and Raghava, G.P. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 65, 40–48

38 Andersen, P.H. *et al.* (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* 15, 2558–2567

39 Tong, J.C. *et al.* (2007) Methods and protocols for predicting immunogenic epitopes. *Brief. Bioinform.* 8, 96–108

40 Zhang, H. *et al.* (2009) Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods. *Bioinformatics* 25, 83–89

41 Hoof, I., et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 61, 1–13.

42 Tenzer, S. *et al.* (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage. TAP transport and MHC class I binding. *Cell. Mol. Life Sci.* 62, 1025–1037

43 Bian, H. and Hammer, J. (2004) Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE. *Methods* 34, 468–475

44 Li, G.B. *et al.* (2004) Predicting allergenic proteins using wavelet transform. *Bioinformatics* 20, 2572–2578

45 Tong, J.C. and Martti, M.T. (2008) Methods and protocols for the assessment of protein allergenicity and cross-reactivity. *Front. Biosci.* 13, 4882–4888

46 Zorzet, A. *et al.* (2002) Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol.* 2, 525–534

47 Centers for Disease Control and Prevention, (2004) Framework for evaluating public heath surveillance systems for early detection of outbreaks; recommendations from the CDC Working Group. *MMWR Recomm. Rep.* 53, RR-5

48 Miller, B. *et al.* (2004) Syndromic surveillance for influenza like illness in ambulatory care network. *Emerg. Infect. Dis.* 10, 1806–1811

49 Prince, A. *et al.* (2005) Containing acute disease outbreak. *Methods Inf. Med.* 44, 603–608

50 Olivio, M. *et al.* (2008) Rule-based knowledge aggregation for large-scale protein sequence analysis of influenza A viruses. *BMC Bioinform.* 9, S7

51 Heiny, A.T. *et al.* (2007) Evolutionarily conserved protein sequences of influenza A virus, avian and human, as vaccine targets. *PLoS ONE* 11, e1190

52 Khan, A.M. *et al.* (2006) Large-scale analysis of antigenic diversity of T-cell epitopes in dengue virus. *BMC Bioinform* 7 (Suppl. 5), S4

53 Peters, H.O. *et al.* (2008) An integrative bioinformatic approach for studying escape mutations in human immunodeficiency virus type 1 gag in the Pumwani Sex Worker Cohort. *J. Virol.* 82, 1980–1992

54 Kaminker, J.S. *et al.* (2007) CanPredict: a computational tool for predicting cancer-associated missense mutations. *Nucleic Acids Res.* 35, W595–W598

55 Hernández, P. *et al.* (2007) Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC Genomics* 8, 185

56 Jonsson, P.F. and Bates, P.A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22, 2291–2297

57 Model, F. *et al.* (2001) Feature selection for DNA methylation based cancer classification. *Bioinformatics* 17, S157–S164

58 Weisenberger, D.J. *et al.* (2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat. Genet.* 38, 787–793

59 Pappalardo, F. *et al.* (2008) Modeling immune system control of atherogenesis. *Bioinformatics* 24, 1715–1721

60 Deisboeck, T.S. *et al.* (2007) Advancing cancer systems biology: introducing the center for the development of a virtual tumor, CViT. *Cancer Inform.* 2, 1–8